Quality Control for Building Libraries from Electrospray Ionization Tandem Mass Spectra

Xiaoyu Yang, Pedatsur Neta, Stephen E. Stein*

Mass Spectrometry Data Center, Mail Stop 8362, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

Abstract

Electrospray ionization (ESI) tandem mass spectrometry coupled with liquid chromatography is a routine technique for identifying and quantifying compounds in complex mixtures. The identification step can be aided by matching acquired tandem mass spectra (MS^2) against reference library spectra as is routine for electron ionization (EI) spectra from gas chromatography/mass spectrometry (GC/MS). However, unlike the latter spectra, ESI MS^2 spectra are likely to originate from various precursor ions for a given target molecule, and may be acquired at varying energies, resolutions, and have characteristic noise signatures, requiring processing methods very different from EI to obtain complete and high quality reference spectra for individual analytes. This paper presents procedures developed for creating a tandem mass spectral library that addresses these factors. Library building begins by acquiring MS^2 spectra for all major MS^1 peaks in an infusion run, followed by assigning MS^2 spectra to clusters and creating a consensus spectrum for each. Intensity-based constraints for cluster membership were developed, as well as peak testing to recognize and eliminate suspect peaks and reduce noise. Consensus spectra were then examined by a human evaluator using a number of criteria, including

fraction of annotated peaks and consistency of spectra for a given ion at different energies. These methods have been developed and used to build a library from >9,000 compounds, yielding 230,000 spectra.

Introduction

Mass spectral reference libraries of electron ionization (EI) spectra are used extensively and routinely to identify compounds separated by gas chromatography [1]. For example, the current NIST/EPA/NIH Mass Spectral EI Library contains spectra for over 200,000 compounds and is a common, tightly-integrated component in many GC/MS data systems [2, 3]. Such use of reference libraries for the identification of electrospray ionization (ESI) tandem mass spectra (MS^2) has, however, been far more limited. While certain MS^2 reference libraries are available for specific applications, such as METLIN [4] for metabolomics, they are often limited to specific platforms or not integrated with instrument data system [4-10]. Also unlike EI libraries [2], documented methods for the quality control of these spectra are not available. Whereas EI libraries developed many years after the field of structural organic mass spectrometry had adopted standard operating practices (e.g., 70 eV ionization at low pressure, precluding ion-molecule interactions), ESI is a relatively new ionization method, without a uniform set of experimental operating conditions. A variety of other features of tandem mass spectra, such as energy dependence, high resolution and multiple precursor ions require new methods for their quality control.

In an effort to extend the scope of reference MS libraries and to document principles useful for other MS^2 data collections, NIST has undertaken the production of a comprehensive ESI MS^2 library for a wide range of molecules [11, 12], intended for use on a variety of platforms and in a range of applications. Different methods are required for development of an ESI MS^2 library in comparison to those used for the odd-electron, positive ion, unit mass resolution MS^1 EI library. ESI MS^2 spectra are the result of even-electron transfer of ionic charge to neutral molecules in solutions at atmospheric pressure, frequently

resulting in simple spectra with sparse fragmentation. The presence of multiple precursor ions and charge states for a single analyte is a necessary consequence of the ESI experiment in which protons or other cations or anions impart charge and form adducts with the neutral species. Additionally, there is a higher degree of spectrum fragment variability due to the energy dependence of the process of generating tandem mass spectra. On the other hand, the high accuracy and resolution available on modern tandem mass analyzers, as well as widely available automated data acquisition and analysis methods, aids in the generation of reproducible, high quality spectra. In common with the EI library, fragmentation pathways are generally not easily predictable and human evaluation of spectra remains an essential step in quality control [2]. This paper presents an overall method developed and tested for building a comprehensive mass spectral library of ESI MS^2 spectra.

EXPERIMENTAL SECTION

Methods

All spectra used in this work were acquired specifically for inclusion in a tandem mass spectral library. They represent a diverse range of compound types and were acquired on a range of instruments, including Collision Induced Dissociation (CID) fragmentation in ion-traps (LTQ) and beam-type collision cells (QQQ, QTOF, and Orbitrap High Energy C-trap Dissociation (HCD) with Fourier transform m/z detection). Infusion of methanol/water/formic acid or acetonitrile/water/formic acid (50:50:0.1) was used for sample introduction and spectra were recorded in either centroid or profile mode. In most cases, the spectra were obtained in the data-dependent mode where MS^2 spectra were acquired for all major ions. Those ions with m/z values corresponding to the precursor types listed in Table 1 were selected for further analysis. Since fragmentation in collision cells is strongly energy dependent, these spectra were acquired over a range of energies from low to high degrees of fragmentation, typically using 10 to 20 collision energies. For some platforms and specific compounds (QQQ, and selected QTOF and Orbitrap measurements), ion selections were made manually for significant precursor ions. When targeted acquisition was used, precursor ions were manually selected among the major precursor ions and could include neutral loss ions arising from in-source fragmentation. In the ion trap measurements, MS^3 and MS^4 spectra were also automatically collected for the top 3 peaks in MS^2 and MS^3 respectively.

Table 1. Precursor types collected for the MS/MS library

Compound Type	ESI Product	Ionic Species
Neutral Molecule	Positive Ions (129,662)	$[M+H]^+$, $[M+2H]^{2+}$, $[2M+H]^+$, $[M+H-H2O]^+$, $[M+H-NH_3]^+$, $[M+H-OH]^+$, $[M+H+H_2O]^+$, $[M+NH_4]^+$, $[M+Na]^+$, $[M-H+2Na]^+$, $[M-2H+3Na]^+$, $[M+K]^+$, $[M-H+2K]^+$, $[M-2H+3K]^+$, $[M+Li]^+$, $[M-H+2Li]^+$, $[M-2H+3Li]^+$
	Negative Ions (23,638)	$[M-H]^-$, $[M-2H]^{2-}$, $[2M-H]^-$, $[M-H-H_2O]^-$, $[M-H-NH_3]^-$, $[M-H+H_2O]^-$, $[M-H+NH_3]^-$
Organic Salt Cations	Positive Ions (1,751)	$[Cat]^+$, $[Cat+H]^{2+}$, $[Cat-H_2O]^+$, $[Cat-NH_3]^+$, $[Cat+H_2O]^+$
	Negative Ions (40)	$[Cat-2H]^-$, $[Cat-2H-H_2O]^-$, $[Cat-2H-NH_3]^-$, $[Cat-2H+H_2O]^-$, $[Cat-2H+NH_3]^-$

*Number in parenthesis is the number of spectra in each category.

Consensus Spectrum Building

Individual library spectra were derived from multiple tandem mass spectral scans acquired for a given precursor ion, at a single energy setting and for a single instrument. No attempt was made to combine spectra from different instruments or different energies. This “consensus” spectrum building, described below, groups similar spectra into the same cluster and creates one consensus spectrum from each cluster. In an effort to reject outliers, consensus spectrum m/z and peak intensity values are taken as the median values of the underlying spectra.

1. Pre-processing. Tandem mass spectral scans with signal/noise ratio (defined as maximum/median peak intensity) ≥ 10 were retained if the difference between the experimental and theoretical m/z values of the precursor ion is less than 0.6 m/z for low resolution spectra and 10 ppm for high resolution spectra.

2. Evaluation of spectral similarity. For each pair of MS/MS spectral scans of the same precursor ion, a dot product value was computed as a measure of spectrum similarity [1].

$$\text{dot product (DP)} = \sum(I_1 I_2)^{1/2} / (\sum I_1 \sum I_2)^{1/2}$$

where I_1 and I_2 are the peak intensities at the same product m/z value. Precursor ion intensities are excluded. For low resolution instruments, the m/z bin size was set at 0.2. For Q-TOF, the m/z bin size was set at 0.02. For high resolution instruments, the m/z bin size was calculated as m/z of fragment ion multiplied by the instrument accuracy (10 ppm), with a low limit of 0.0040 m/z (for fragment ions with m/z < 400). For example, for product ion with m/z 700, the bin size is 0.0070. Note that DP is 1 for identical spectra and 0 for spectra with no peaks in common.

DP provided a robust overall level of spectrum similarity, and worked well for spectra with multiple major peaks. However, for spectra with few major peaks, where less abundant peaks were more diagnostic, the DP measure was inadequate to reject outlier spectra. A special filter was applied for this purpose, namely, if any of the peak intensity ratios in a pair of spectral scans exceeded a specified threshold ratio, spectra would be placed in different clusters. This threshold ratio depended on the intensity of the larger of the two peaks being compared, as follows: When the greater of two peaks was higher than 75% of the largest (base peak) the threshold value was 3; when the larger fell between 50% and 75% the threshold was 4, for larger peak from 25% to 49% the threshold was 5. No restrictions were placed on pairs of peaks both less than 25%.

3. Clustering.

(a) DP was calculated for each pair of spectra.

(b) The spectral scan with the greatest number of matches (DP > 0.7) was assigned as the starting spectrum or cluster seed. All spectra with DP > 0.7 matching the seed were assigned to that cluster.

(c) Step (b) was applied to the un-clustered spectra to find spectra in the next most populated cluster. This was repeated until no un-clustered spectra remained.

(d) All of the spectra, except for the seed of each cluster, were then subjected to iterative cluster reassignment. This was done by assigning each spectrum to the cluster whose seed was closest (highest DP) and satisfied the peak ratio requirement described above.

4. Creating "consensus" spectra. This process creates a single "consensus" spectrum from each cluster of spectra created by step 3. Peaks from each scan in the same cluster were collected to form a 'composite' spectrum. Beginning with the most intense peak in the 'composite' spectrum (if there was more than one most intense peak, the median m/z of these peaks was selected), all the peaks within an m/z bin (bin size defined in Section 2) were initially placed in the same bin in the consensus spectrum. Peaks in each bin were then converted to a single consensus peak by taking the medians of the m/z and abundance values. This procedure was repeated for bins containing the next highest peak until all peaks were included in their respective bins for the consensus spectrum.

Depending on acquisition conditions, consensus spectra could contain from 5-500 individual spectral scans. MS^3 and MS^4 spectra were processed in the same way as MS^2 spectra.

Quality Control

The quality of the consensus spectra generated by the above procedures was further improved by the following steps.

1. Noise Removal: Satellite peaks nearby major peaks that were apparently generated by imperfect centroiding were removed as described in Supplement A. Other varieties of noise peaks were removed by peak "voting". Any m/z peak in a consensus spectrum that was present in fewer than 25% of the individual spectra or appeared in only one spectrum was removed. This primarily eliminated peaks whose intensity was less than 0.5% of the base peak. This cutoff level of 25% was determined from analysis of 6,723 high-resolution HCD consensus spectra as follows. The percent occurrence of each peak in a consensus spectrum among the spectra used to create it was first computed. Then, at each occurrence level, we calculated the fraction (*f*) of peaks that could be assigned a chemical formula

consistent with the precursor formula. From the plot of *f* vs. occurrence (Figure 1, solid circles) we chose a cutoff level at fraction *f* = 0.5, (50% of peaks at that level can be assigned a chemical formula). This led to the removal of peaks that appear in less than 25% of the individual spectra. We found that 98% of the peaks removed by this method were of low intensity, < 0.5% of the base peak (open circles) and < 0.001% of the peaks had intensity > 10%. Note that this process made no distinction between peaks that could or could not be assigned chemical formulas. Selectively removing peaks without plausible formulas would have, in effect, removed evidence of contamination, whose presence could be helpful in later evaluations. Since along with each peak in a library spectrum is given the most probable chemical formula, search algorithms can readily ignore these peaks if desired, so their presence need not have an effect on library search scoring.

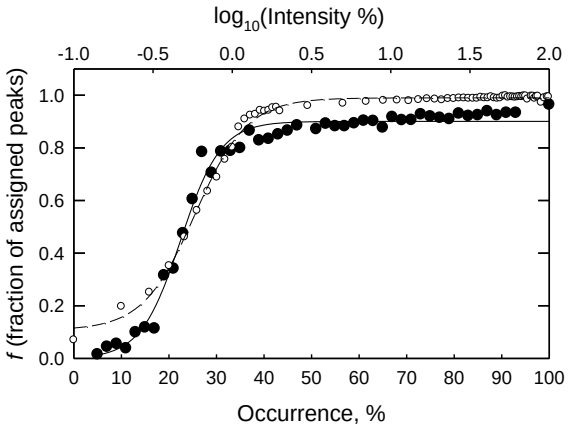

Figure 1. Plots of *f* (fraction of formula-assigned peaks) vs. occurrence level (solid circles) and $\log_{10}$ Intensity (open circles) for ~ 7 million peaks of 6,723 high-resolution Orbitrap HCD spectra.

2. Peak Annotation: This procedure attempted to identify and label each fragment peak in order to judge spectrum quality. For peptide ion spectra, peaks were annotated using conventional notation (precursor, y, b, a, common neutral loss, immonium, and internal fragment ions). For high resolution spectra of non-peptide ions, the peaks were annotated with the most probable chemical formula

consistent with the precursor formula [13] and mass accuracy. The accuracy of assignment of each peak in the spectrum was calculated as follows:

$$\text{Accuracy (ppm)} = \frac{|(\text{Observed m/z} - \text{Theoretical m/z})|}{\text{Observed m/z}} \text{ x } 10^6$$

Any high resolution peak that did not match the closest possible fragment formula within 10 ppm was labeled as unassigned. Peaks of minor isotopes were assigned only if the corresponding major isotopic peaks were found and if the intensity ratio of the two peaks was within a factor of 10 with the calculated natural abundance ratio of isotopes. A measure of spectrum quality was the percentage of unassigned MS^2 peak intensity:

$$\text{Unassigned (\%)} = \frac{\text{Sum of intensities of unassigned peaks}}{\text{Sum of intensities of all peaks}} \text{ x } 100$$

If this value was > 10%, the spectrum was flagged for deletion or re-measurement.

3. Collision energy dependence: Ions that were analyzed at different collision energies are expected to follow a well-defined trend, i.e. the precursor peak intensity decreases monotonically with increasing energy and fragment ion intensities increase with energy as they are produced and then decrease at higher energies as they fragment further. Moreover, the average masses of peaks monotonically declines with increasing energy. This behavior was required for acceptance of spectra for the library. However, this criterion could fail if spectra from different runs were intermixed, due to uncontrollable differences in collision conditions. In cases where several runs were made, all spectra were derived from the best single run after manual inspection.

The overall procedure of building a tandem mass spectral library is summarized in Figure 2. This figure includes the procedures described above, as well as other quality control procedures such as consistency of the chemical information for each compound, which are summarized in the Supplement B.

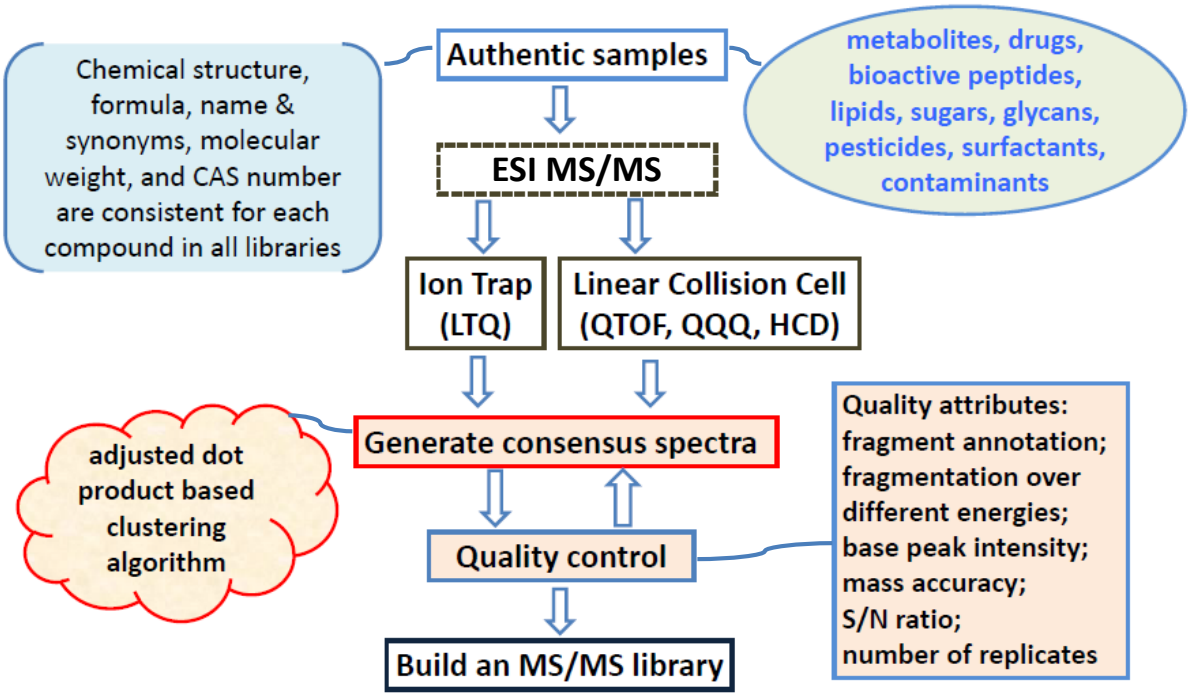

Figure 2. Procedure of building a tandem mass spectral library

Results and Discussion

The procedures described in the Methods section were tested using spectra obtained for >1,000 compounds. The following sections present a few examples of the way these procedures were utilized.

Creating consensus spectra:

The idea of using consensus spectra for improving the quality of tandem spectra has been shown for peptide spectra and continues to be employed [14, 15]. MS/MS measurements made for building the present library usually generate many (>10) replicate spectra of each precursor ion. Some of these may differ due to unstable instrument conditions, impurity peaks, or carryover from previous samples. Spectral clustering accepts only the most repeatable spectra for consensus spectrum creation. Combining these spectra into a single consensus spectrum leads to rejection of spurious peaks and to higher quality library spectra.

Peak intensity is a critical factor in evaluating spectral similarity. We observed that when the number of peaks is small, especially for small molecules, the dot product value might not represent the intuitive similarity of a pair of spectra. To overcome this problem, we used peak ratio criteria as an additional restriction when creating clusters (see Methods). This method was tested with > 10,000 MS/MS spectra of known compounds on LTQ, QTOF, and Orbitrap HCD instruments, and validated by manually inspecting consensus versus original spectra. This algorithm assists in the identification and deletion of low quality spectra, spectra of contaminants with similar precursor m/z, and spectra with peaks from co-fragmented contaminants. Depending on compound structure and experimental condition, 1-10 clusters were generated for each precursor ion. Of the spectra included in the library, 94.6% had only one cluster, 4.7% had two clusters, and 0.7% had more than two clusters.

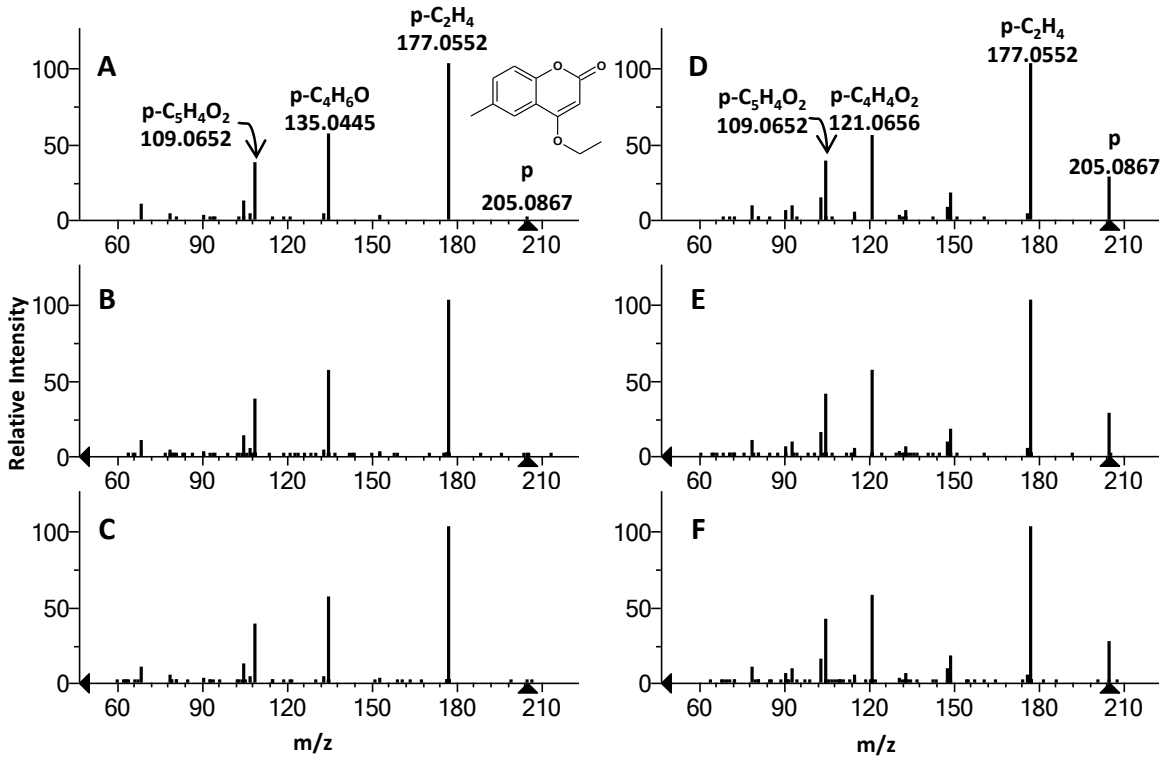

Figure 3. An example of clustering of 29 spectra acquired on HCD at collision energy 26 eV for the $[M+H]^+$ ion of 4-ethoxy-6-methylcoumarin and generation of a consensus spectrum from each cluster.

Two clusters of 21 and 8 spectra were formed. A is the consensus spectrum generated from cluster 1. B and C are 2 individual spectra out of 21 spectra from cluster 1. D is the consensus spectrum generated from cluster 2. E and F are 2 individual spectra out of 8 spectra from cluster 2. Letter 'p' stands for the precursor of the compound. In this example, p is $[M+H]^+$. The same annotation is used throughout the library.

Figure 3 shows an example where 29 Orbitrap HCD spectra of the same precursor ion, acquired at the same collision energy, are divided into different clusters and a consensus spectrum was generated from each using the clustering algorithm. The main difference between the two clusters is intensities of peaks at m/z 121.0656 and 135.0445. After evaluation, consensus spectrum A was preferred to D because the intensity of the highest peak is higher (by a factor of >300) and the number of spectra included is larger (21 vs. 8). Therefore, consensus spectrum A was chosen for the library. Furthermore, the spectra for generating consensus spectrum D appeared at acquisition times < 8 min, where the signals had low intensities, whereas the spectra for consensus spectrum A appeared at > 8 min and had much higher intensities. This suggested that the low intensity spectra for consensus spectrum D are due to carryover. In fact, the peak at m/z 121.0656 was a major peak in the prior sample. Although the injection system is routinely washed between samples, and in most cases we do not detect any carryover, this example shows that our clustering method allows us to eliminate the effect of carryover when it does happen. Such experimental factors are one of the principal reasons for requiring human evaluation of all spectra prior to inclusion in the final library.

Figure 4 illustrates another example of clustering 14 ion-trap spectra of the [M+H]+ ion of chromone-2-carboxylic acid. Spectrum B was eliminated because of its dissimilarity with the other 13 spectra (2 of which are shown in the figure).

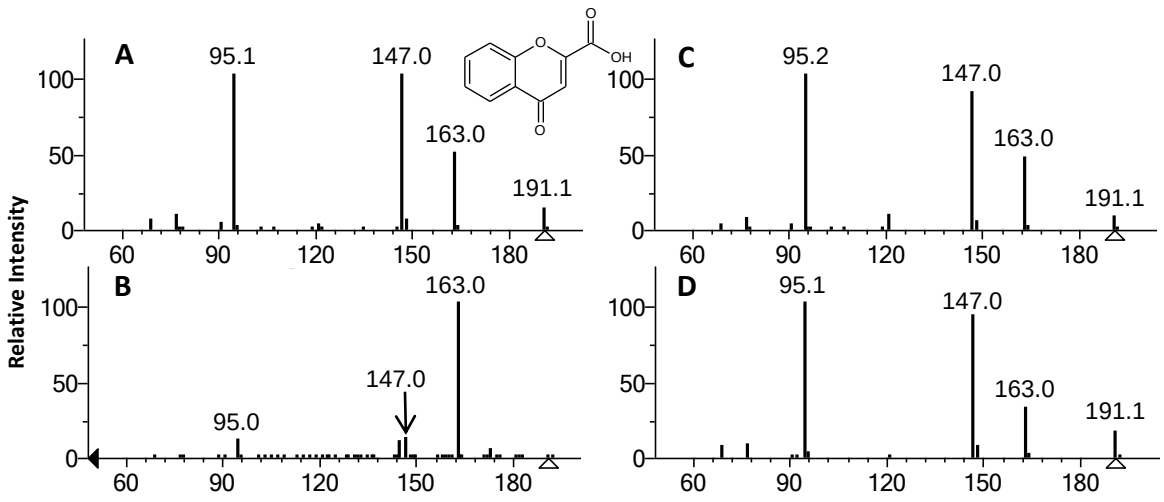

Figure 4. An example of clustering of 14 spectra acquired on an ion-trap mass spectrometer for the $[M+H]^+$ ion of chromone-2-carboxylic acid. Only one cluster from 13 spectra was generated. A is the consensus spectrum generated from the cluster. B is the spectrum that is excluded from the cluster. C and D are 2 individual spectral scans out of 13.

A general problem with any clustering method for high resolution spectra arises from splitting and merging of peaks, which can depend on signal strength as well as instrument resolution. In the present approach such problems generally placed spectra with major split/merged peak differences in different clusters - the one with the largest membership was then selected for the library.

Noise Removal:

Comparison of individual MS/MS spectra collected for the same ion under the same conditions generally exhibit variations in peak intensities and in exact m/z values, and certain peaks may appear in some spectra but not others [16-18]. These random fluctuations are generally removed by averaging a number of spectra. Individual peaks which occur more frequently are more likely to be true peaks, while low frequency peaks may be noise. A voting algorithm was applied to reject these infrequent peaks. It is also observed that for most of the spectra, the intensities of infrequent peaks were generally very low (< 0.5 % of the base peak intensity). An example is given in Figure 5, which shows the noise removal results for one consensus spectrum by using this voting algorithm. Table S1 in Supplement A shows that all the

peaks that are eliminated occur less frequently and have low intensities. Figure S1 and S2 in Supplement A show examples of removing noise peaks around a high intensity peak due to apparent centroiding issues from Orbitrap HCD and QTOF spectra, respectively.

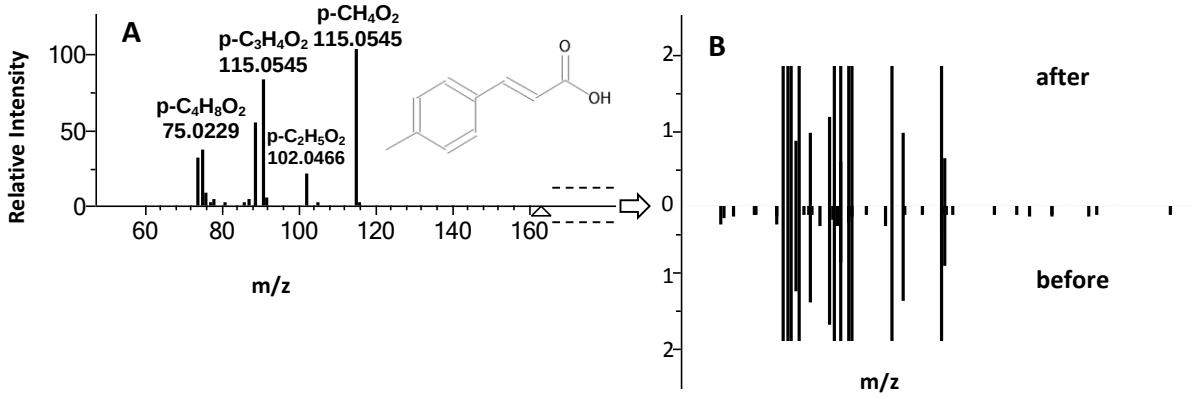

Figure 5. An example of noise removal in a consensus MS/MS spectrum for the 4-methylcinnamic acid $[M+H]^+$ ion. A: The consensus spectrum with 16 peaks was generated from 12 spectra after noise removal. B: Head to tail comparison of spectra before and after noise removal. The accepted consensus ions are shown in the top spectrum; the unmatched peaks in the bottom spectrum were either noise or contaminant/impurity peaks.

Quality control:

Consensus spectra for most ions were generated under different collision energies. The general fragmentation pattern can be utilized to eliminate low quality spectra when the fragmentation in one spectrum is not consistent with those at higher and lower energies. For example, the spectrum at collision energy 5V in Figure 6 was removed due to the absence of a peak at m/z 88.9.

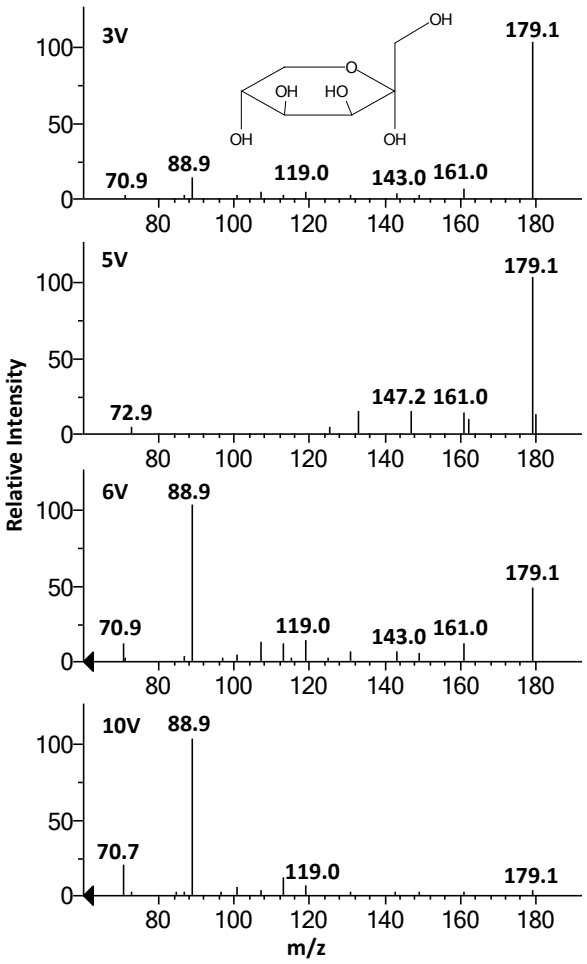

Figure 6. An example of correlating and assessing peak intensities with respect to collision energies to omit uncharacteristic spectra for the $[M-H]^-$ ion of D-(-)Tagatose acquired on QQQ. The spectrum at 5 V was eliminated since m/z 88.9 was absent from this spectrum.

Spectral accuracy and peak annotation

For high accuracy Orbitrap mass spectra, we attempted to assign a chemical formula to each fragment ion, consistent with the formula of the precursor ion, with accuracy within 10 ppm. If a peak with intensity > 20% (compared to the base peak) remained unassigned, or the total intensities of all unassigned peaks was > 10%, the spectrum was rejected and the sample was rerun. We tested this procedure with 1994 spectra for 88 compounds. When the mass accuracy of the peak assignment was limited to ≤ 10 ppm, 1893 (92%) spectra contained ≤ 10% unassigned total peak intensities. When the

mass accuracy limit was 7, 6, and 5 ppm the number of spectra with ≤ 10% unassigned peak intensity decreased (Table 2).

Table 2. Peak annotation for 1994 spectra from HCD and FTCID of 88 compounds using different accuracy levels

Accuracy (≤ ppm)	N
10	1893
7	1729
6	1654
5	1492

N is the number of spectra containing ≤ 10 % unassigned total peak intensity.

For fragment ions with higher masses a number of chemical formulas are possible, even using the stringent allowance of 1 ppm [19]. We compared the results with the 1994 spectra mentioned above, using ≤ 5 or ≤ 10 ppm accuracy. With ≤ 5 ppm, we found only two cases with ions having multiple formula assignments, and their intensities were < 1%. With ≤ 10 ppm accuracy, the number of peaks with more than one assignment was higher, ≤ 1 % of the peaks, with a median intensity < 1 %. These results indicate that formula annotation with ≤ 5 ppm mass accuracy achieves high spectral quality, but ≤ 10 ppm accuracy produced satisfactory results for 97 % of spectra. Using 5ppm accuracy did not improve the spectrum quality compared with 10ppm, and consequently, 10ppm was used for peak annotation.

The presence of contaminant ions with the same m/z is not a general problem in our measurements because individual, purified compounds are employed. The precursor purity was ascertained by the lack

of extraneous peaks in the MS^1 spectrum within the fragmentation window of the precursor ion (typically ± 1.5 m/z), and by the relative intensities of the isotopic peaks. Precursor purity, computed using in-house software, was generally ≥ 95% and considered in all evaluations. Spectra with lower precursor purity were generally rejected.

MS/MS spectra under different conditions and MS^n

Spectra for most compounds were acquired using both ion trap and variable energy (beam-type) collision cell measurements (see examples in Supplementary Figure S3). This permits matching of CID query spectra from most modern instruments.

In many measurements with the LTQ and the Orbitrap, MS^n results were also collected, to help with further confirmation of ion trap identifications. An example of MS^n spectra from the Orbitrap ion trap (Supplementary Figure S4) shows successive fragmentation steps in MS^3 and MS^4 and helps identify the fragmentation pathway.

Peptides

The peptide component of this library consists mainly of bioactive peptides and di- and selected tri-peptides. The bioactive peptides were obtained from commercial sources and included hormones, neuropeptides, alkaloids, antibiotics, etc. The di- and tri-peptides include all possible 400 dipeptides and 800 tryptic tripeptides (with arginine or lysine at the C-terminus), and are common constituents of biological fluids and protein digests, yet because of the small number of sequence product ions, cannot be identified unambiguously by the sequence search engines commonly used in proteomics.

The ions of each peptide in the spectrum were annotated as *p* (precursor), *y*, *b*, *a* fragment ions, isotopic ions, ions formed by loss of water or ammonia from precursor and fragment ions, immonium ions, and other known specific fragments. These MS/MS spectra are expected to find use for both metabolite identification and proteomics. For example, in Figure S5 in Supplement C, we compared the spectra of the tripeptides AFK and FAK in an attempt to distinguish between these similar sequences.

The figure shows that the fragment ion y2 from each spectrum is the signature peak to distinguish these peptides from each other.

Critical inspection of spectra

All the consensus spectra were manually inspected by an experienced mass spectrometrist. Long experience with building evaluated spectral libraries has shown us that human inspection is an essential component[2]. Since the spectra were recorded in a data-dependent mode, various precursor types were detected for the same compound (see Table 1) and many consensus spectra were generated for each compound. About 55% of these spectra are of low quality and are easily removed by human inspection. The remaining spectra were examined in more detail to confirm that all major peaks are plausible fragmentation products for the precursor ion. When relevant, the expected correlation of peak intensities with collision energy was also used in this process[20]. Additional 14% of the spectra were filtered out by this inspection.

Peaks in high resolution MS/MS spectra were annotated with the most probable chemical formula consistent with the precursor formula. When more than one formula is possible, the formulas are listed in order of m/z proximity to the ion. In order to minimize the number of close assignments, we retained at least one carbon in the formula. Following this procedure a number of fragment ions that do not contain carbon may remain unassigned. A table of 64 such non-carbon containing ions was developed during spectrum evaluation and used to assign peaks automatically when their composition was consistent with that of the precursor (see Supplement D).

A total of 234,329 spectra of 9,345 compounds from various instruments comprise the current library. Of these,187,733 are positive ions (92,632 are $[M+H]^+$), whereas 46,596 are negative ions (20,052 are $[M-H]^-$). Detailed precursor types and the numbers of spectra are listed in Table 1. The utility of including all major precursor ions for certain compounds is illustrated in Figure 7 for a selected example, inosine 5'-diphosphate, where four positive ions and two negative ions, each having rich, characteristic spectra,

were acquired for this compound. The number of precursor ions is 41,849 and 14,836 respectively for ion-trap and collision cell (QTOF, QQQ, and HCD). This library collection appears unique in its broad coverage, inclusion of alternative precursor ions and representation of ion-trap and beam-type collision-cell fragmentation, as well as its extensive use of computer-assisted evaluation. An earlier version of this tandem mass spectral library has been successfully applied in the metabolite identifications [21, 22].

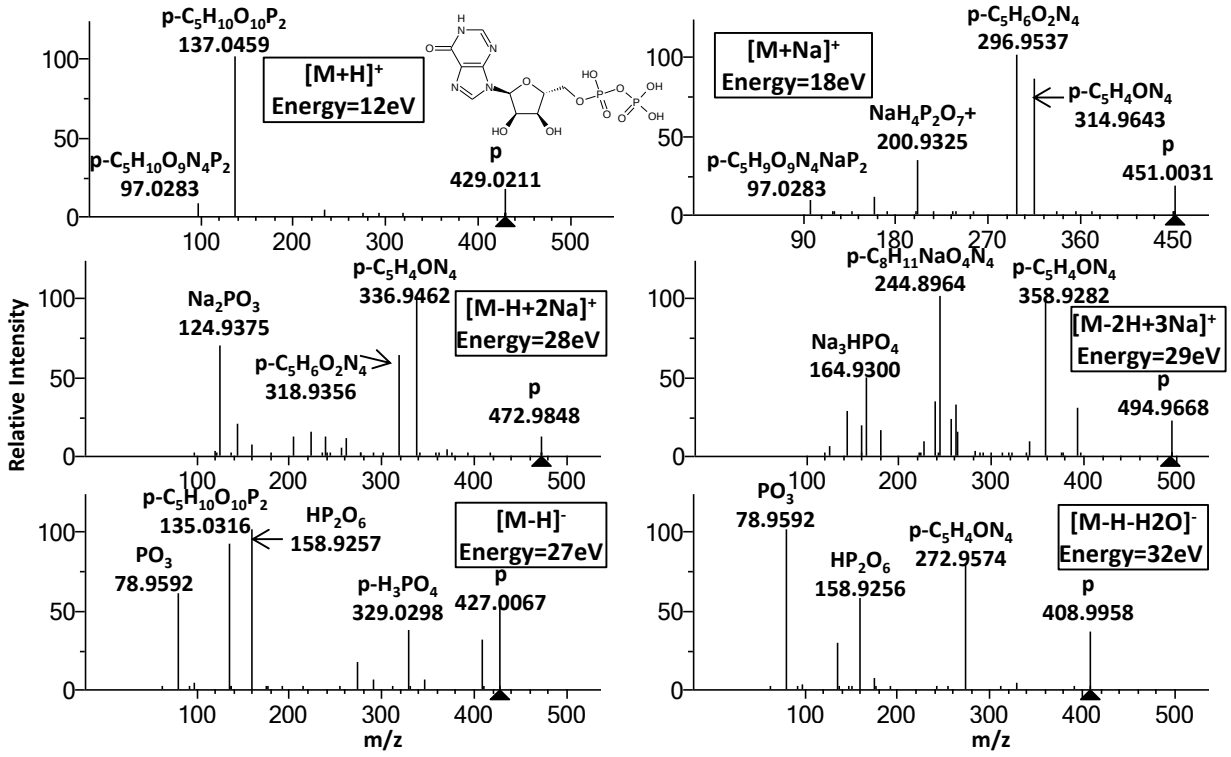

Figure 7. An example of various precursor types of inosine 5'-diphosphate acquired on Orbitrap HCD.

Conclusion

A clustering algorithm, using an intensity-restricted dot product measure of spectral similarity, along with various filtering methods, such as peak 'voting' and annotation, has been shown to provide a robust means of creating information-rich, reference quality consensus spectra of a broad range of individual compounds. For wide use, spectra were acquired in both multistage ion-trap instruments and variable energy 'beam-type' collision cells, with processing methods taking explicit account of

instrument resolution. When acquired in a single run, variations of peak intensities with collision energies served to provide additional quality control. These spectra were individually examined by an experienced mass spectrometrist to remove low quality and erroneous spectra and to recommend re-measurement. The resulting comprehensive library, in both compound class and instrument type, are intended to serve much the same purposes for tandem mass spectrometry as electron ionization libraries do today in GC/MS analysis.

ASSOCIATED CONTENT

Supporting Information

This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Email: xiaoyu.yang@nist.gov. Phone: +1-301-975-2501.

Acknowledgments

We thank Drs. Karl K. Irikura and Sanford Markey for detailed discussion and critical review of this manuscript. We also thank Drs. Yuri Mirokhin, Dmitrii Tchekhovskoi, Quan-Long Pu, Lisa Kilpatrick, Mark Wingerd, Yufang Zheng, Yamil Simon, Yuxue Liang, Maria Lorna A. De Leoz and Wei Mi for discussions on data analysis, instrument settings and collecting some of the data for the algorithm testing.

References

(1) Stein, S.E. J. Am. Soc. Mass Spectrom. **1995**, 6, 644-655.

(2) Ausloos, P.; Clifton, C.L.; Lias, S.G.; Mikaya, A.I.; Stein, S.E.; Tchekhovskoi, D.V.; Sparkman, O.D.; Zaikin, V.; Zhu, D. *J. Am. Soc. Mass Spectrom.* **1999**, 10, 287-299.

(3) NIST/EPA/NIH Mass Spectral Library. http://www.nist.gov/srd/nist1a.cfm.

(4) METLIN. http://metlin.scripps.edu/.

(5) Massbank. http://www.massbank.jp/?lang=en.

(6) Wishart, D.S.; Jewison, T.; Guo, A.C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y., Mandal, R., Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. *Nucleic Acids Research* **2013**, 41, D801-807.

(7) MSforID. http://msforid.com/.

(8) Dresen, S.; Gergov, M.; Politi, L.; Halter, C.; Weinmann, W. *Anal. Bioanal. Chem.* **2009**, 395, 2521-2526.

(9) Broecker, S.; Herre, S.; Wust, B.; Zweigenbaum, J.; Pragst, F. *Anal. Bioanal. Chem.* **2011**, 400, 101-117.

(10) Go, E.P. *Journal of Neuroimmune Pharmacology* **2010**, 5, 18-30.

(11) NIST/EPA/NIH Mass Spectral Library. http://chemdata.nist.gov/mass-spc/msms-search.

(12) Stein, S. *Anal. Chem.* **2012**, 84, 7274-7282.

(13) Kind, T.; Fiehn, O. *BMC Bioinformatics* **2007**, 8, 105 doi:10.1186/1471-2105-8-105.

(14) Lam, H.; Deutsch, E.W.; Eddes, J.S.; Eng, J.K.; Stein, S.E.; Aebersold, R. *Nature Methods* **2008,** 5, 873-875.

(15) Lam, H.; Deutsch, E.W.; Eddes, J.S.; Eng, J.K.; King, N.; Stein, S.E.; Aebersold, R. *Proteomics* **2007,** 7, 655-667.

(16) Hassell, K.M.; LeBlanc, Y.C.; McLuckey, S.A. *Anal. Chem.* **2011**, 83, 3252-3255.

(17) Busch, K.L. *Spectroscopy* **2002**, 17, 32-34.

(18) Du, P.C.; Stolovitzky, G.; Horvatovich, P.; Bischoff, R.; Lim, J.; Suits, F. *Bioinformatics* **2008**, 24, 1070-1077.

(19) Kind, T.; Fiehn, O. *BMC Bioinformatics* **2006**, 7, 234. doi:10.1186/1471-2105-7-23.

(20) Neta, P.; Simon-Manso, Y., Yang, X., Stein, S.E. *J. Am. Soc. Mass Spectrom.* **2009**, 20, 469-476.

(21) Lowenthal, M.S.; Phillips, M.M.; Rimmer, C.A.; Rudnick, P.A.; Simon-Manso, Y.; Stein, S.E.; Tchekhovskoi, D.; Phinney, K.W. Anal. *Bioanal. Chem.* **2013**, 405, 4451-4465.

(22) Simon-Manso, Y.; Lowenthal, M.S.; Kilpatrick, L.E.; Sampson, M.L.; Telu, K.H.; Rudnick, P.A.; Mallard, W.G.; Bearden, D.W.; Schock, T.B.; Tchekhovskoi, D.V.; Blonder, N.; Yan, X.; Liang, Y.; Zheng, Y.; Wallace, W.E.; Neta, P.; Phinney, K.W.; Remaley, A.T.; Stein, S.E. *Anal. Chem.* **2013**, 85, 11725-31.